L&PS – Logic and Philosophy of Science
Vol. IX, No. 1, 2011, pp. 405-411

Evolutionary Representations of Biological History

Ludovica Lorusso
Dipartimento di Scienze Politiche, Scienze della Comunicazione e Ingegneria dell'Informazione
Università di Sassari
e-mail: lorusso@uniss.it

ABSTRACT. Phylogenetic systematics is the branch of biology that reconstructs the history of biological entities. There are many possible representations of such history, like trees and networks, and different histories for different biological levels, like histories of species and genes. Since Darwin a unique tree named "Tree of Life" has been considered as the best representation of the history of species. In this paper I would like to question whether the tree of life is the best representation of the historical relationships among species in the light of the evolutionary theory and the biological evidence.

1. Introduction

In *The Origin of Species* Darwin (1859) established the link between what he called a "natural" classification of species and the "universal tree of life": any classification to be natural must come from the universal tree, in which each species occupies a unique position that represents exactly its position in the history of species.

Trees are certainly the most famous representations of the history of relationships among biological entities. A tree is a reliable representation of history only under the assumption that the evolutionary process is a hierarchical and branching process (Mayr 1982). In a hierarchy each species is part of one and only one genus, each genus is part of one and only one family, and so forth; in a branching process an ancestral species A originates two daughter

Published by L&PS – Logic and Philosophy of Science
http://www2.units.it/ ~episteme — ISSN: 1826-1043

species, B and C and these two species could have, in turn, some offspring or none (Doolittle 1999). To make a historical inference in biology means to reconstruct a phylogeny at different biological levels like genes, populations, species and higher taxonomic groups. Since any phylogeny is thought to have been developed in an evolutionary way, the terms evolution and phylogeny of species tend to be interchangeable; unfortunately, we will see that there are evolutionary processes that are dismissed inside models used in phylogenetic methods.

In phylogenetic methods a classification is needed to reconstruct trees and therefore such classification can not come from the tree. However, such classification should follow a genealogical concept of species in order to generate a reliable reconstruction of the history of species (see e.g., Velasco 2008) and the genealogical concept of species is based on a previous reconstruction of historical relationships between species, for instance, a tree. This may not be a real circularity, because the previous tree may be reconstructed on the basis of a different kind of evidence (for example, phenotypic characters), while current methods in molecular phylogenetics use molecular evidence in order to reconstruct trees. Anyway, the relation between trees and classifications is not simple or one-way. Finally, there are two main problems with trees and classifications: first, classifications based on different concepts of species lead to different phylogenetic trees; second, trees reconstructed from different evidence (i.e., phenotypic and molecular characters) are different, leading to different genealogically based classifications. It seems obvious that both these issues represent a serious obstacle to the Darwinian realist concept of "natural" classification: first of all, a classification needs to be settled before reconstructing a tree; second, how to decide for *the* "natural" classification among many classifications? Also, these issues constitute an obstacle to the aim of phylogenetics to reconstruct the tree of life. However, should the existence of many trees be considered a failure of phylogenetic methods in representing the "true" history of species or otherwise all these trees need to be considered useful epistemological tools in biology? In order to understand the problem in the next paragraph I will introduce the current methods in phylogenetics based on molecular data.

2. Species trees from molecular data

Nowadays molecular data are mostly used to reconstruct species trees. Molecular evolution or phylogenetics is the branch of biology that reconstructs evolutionary histories of species and genes by using molecular data, like DNA sequences. Not all DNA sequences can be used to reconstruct trees, but only "homologous" sequences, which are the sequences shared by two taxa because inherited from a common ancestor. Any tree reconstructed from homologous sequences of DNA is based on the *coalescent theory*. A "coalescent event" occurs when two lineages of DNA molecules merge back into a single DNA molecule at some time in the past. Hence, a coalescent event is the time inverse of a DNA replication event. By the random process of genetic drift, some molecules get more copies into the next generation than others. This process causes fixation of some molecules and extinction of others. All of the copies of a homologous stretch of DNA at the present time can be traced back in time to a common ancestral DNA from which all current copies are descended. All these DNA sequences related by a common ancestry share a gene tree history, in which nodes refer to cases of DNA replication. In phylogenetic methods gene trees are often represented by haplotype trees, where a haplotype tree is a gene tree in which mutation occurred within one branch after replication, making evident the coalescent event. Given a sample of haplotypes that arose solely from mutations, an evolutionary tree of the haplotypes exists that describes the history of mutational accumulation in DNA lineages that yield the current array of haplotype variation.

Two different kinds of trees can be reconstructed from DNA sequences: species trees and gene trees. While gene trees represent the history of a certain sequence of DNA, species trees based on DNA sequences represent either a *consensus tree* reconstructed from multiple-sequences alignment (Delsuc et al. 2005; Gadagkar et al. 2005) or a *concordance tree* built from clades coming from different gene trees and shared by a plurality of the genome (see e.g., Baum 2007). Both consensus and concordance trees aim to be the unique "true" tree of life, reflecting the real history of species.

The problem is that phylogenies of species are often different from phylogenies of genes: because DNA lineages can be carried across speciation events, coalescence times are sometimes older than the species and therefore in these cases gene trees are *not concordant* with species trees. Moreover, different genes have different histories within a population and within a species: any part of the genome has its own history, mostly independent of the histories of the other parts (see e.g., Maddison 1997). For this reason a partial selection of sequences used to reconstruct phylogenies of organisms can lead to

trees that fail in reflecting the history of those organisms. Only if the homologous sequences are "orthologous", a gene tree will reflect a species tree. Homologous sequences are orthologous if they were separated by a speciation event: when a species diverges into two separate species, the divergent copies of a single gene in the resulting species are said to be orthologous. Sequences generated by a duplication event are named "paralogous" (see Figure 1).

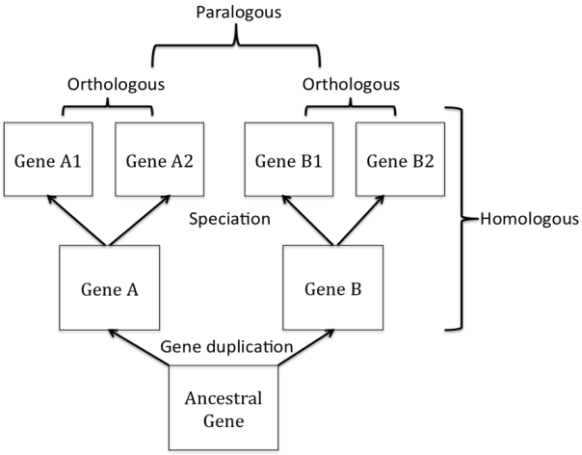

Figure 1. The orthology and paralogy relationships between genes. Genes A and B are paralogous and so are their descendents. Genes A1 and A2 are orthologous and so are genes B1 and B2. All members of the gene families A and B are homologous.

In current phylogenetic methods the orthology assessment is based on the "total evidence principle", which consists in a choice of an arbitrary percentage of similarity between DNA sequences. The use of such principle in selecting sequences does not consider the evolutionary significance of similarities. The problem is that many evolutionary events like differences in evolutionary rates and base composition, and the occurrence of horizontal gene transfer (HGT) among species may generate non orthologous similarities. For this reason orthology assessment should not be based on an arbitrary percentage of similarity; instead, it should require a rigorous phylogenetic analysis of individual genes in order to know the evolutionary processes that have shaped their variation. There is another serious methodological problem in reconstructing trees from DNA sequences: all phylogenetic methods assume models that are not consistent with the theory of evolution, like for example models without natural selection. Another evolutionary process dismissed in these

methods is the HGT, which is known to be common among many species; similarities generated through this process over the years have created horizontal relationships among species. Because such relationships are not hierarchical but reticulated, a tree can not be the right representation of them and other non tree-like representations are needed, for instance, networks.

The use of models that dismiss processes known to have played an important role in shaping the current biological variation among species is a real thorn in the side of phylogenetics.

3. The epistemological framework: Many trees are better than one

In phylogenetics there is a large debate on the fact that both consensus and concordant trees seem to be inconsistent with the biological evidence (see e.g., Delsuc et al. 2005).

This debate can be solved by choosing a non realist framework where the goal of phylogenetics is no more to reconstruct the unique "universal tree". If different genes have different histories that are better represented by different trees and networks, why should phylogenetics insist in reconstructing a unique tree of species based on genes? The main purpose of phylogenetics should be to investigate the evolutionary information inside genes and the relationships between this information and the history of organisms and species: "The reconstruction of the topology of the organismal phylogeny is not in itself the ultimate goal. The challenge is to understand the evolutionary history of organisms and their genomes, the functions of their genes, and how this relates to their interaction with the environment." (Delsuc et al. 2005)

The problem of using models that are not consistent with the theory of evolution and the biological evidence is not so serious as long as we are able to compare trees reconstructed from different evidence (i.e., phenotypic characters, different genes), because by comparing different trees it is possible to control the effects that these models have on the historical relationships among biological entities. The tree of life does not allow to control such effects and for this reason it is not a testable representation of the evolutionary past. Moreover, only if we do not lose the biological information coming from different parts of genomes and different phenotypic characters it is possible to create the best evolutionary models; for example, we should not use the assumption of the absence of natural selection when using sequences or phenotypic characters that are co-evolved by natural selection. Different evolutionary processes may have acted on different parts of the genome and therefore

similarities across different parts of the genome have different evolutionary significances; if sequences are compared with the criterion of a percentage of similarity, all the biological information that different parts of genomes convey gets lost and finally we will obtain a unique tree with no information.

Consider for example human populations that share different sequences for different evolutionary reasons. In a "Mendelian" population, individuals share genetic variations because this population represents a reproductively isolated community that preserves a specific gene pool; however, two individuals from different Mendelian populations can share specific genetic variations because of the fact that they are co-evolved by natural selection (see e.g., Templeton 2005). While the Mendelian gene pool has been shaped through a history of inbreeding, specific genetic variations have been generated by a process of natural selection acting on organisms living in similar environmental conditions.

The history of populations and species can be represented by different trees and networks that tell us the different ways in which evolutionary processes have been shaping biological variation. Phylogenetics should not aim to reach a unique tree of life, but it should aim to understand the evolutionary history of variability among organisms; if this aim can be better reached by means of reconstructing many trees or networks, this is what phylogenetic methods should do.

REFERENCES

BAUM, D. A. (2007): "Concordance trees, concordance factors, and the exploration of reticulate genealogy", *Taxon*, 56, pp. 417-426.

DARWIN, C. (1859): *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, London: John Murray.

DELSUC, F., BRINKMANN, H., HERVÉ, P. (2005): "Phylogenomics and the reconstruction of the tree of life", *Nature Reviews, Genetics*, 6, pp. 361-375.

DOOLITTLE, W. F. (1999): "Phylogenetic Classification and the Universal Tree", *Science*, 284, pp. 2124-2128.

GADAGKAR, S. R., ROSEMBERG, M. S., and KUMAR, S. (2005): "Inferring Species Phylogeneies from Multiple Genes: Concatenated Sequence Tree versus Consensus Gene Tree", *Journal of Experimental Zoology (Mol Dev Evol)*, 304B, pp. 64-74.

MADDISON W. P. (1997): "Gene Trees in Species Trees", *Systematic Biology*, 46, pp. 523-536.

MAYR, E. (1982): *The Growth of Biological Thought*, Cambridge: Belknap.

TEMPLETON, A. R. (2005): "Haplotype Trees and Modern Human Origins", *Yearbook of Phisical Anthropology*, 48, pp. 33-59.

VELASCO, J. (2008): "Species concepts should not conflict with evolutionary history, but often do", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 39, pp. 407-414.